Mathématiques et sciences humaines
Mathematics and social sciences

173 | Printemps 2006
Varia

Colinéarité et régression linéaire

Collinearity and linear regression analysis

Thierry Foucart

Édition électronique
URL : http://journals.openedition.org/msh/2963
DOI : 10.4000/msh.2963
ISSN : 1950-6821

Éditeur
Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée
Date de publication : 1 mars 2006
ISSN : 0987-6936

Référence électronique
Thierry Foucart, « Colinéarité et régression linéaire », *Mathématiques et sciences humaines* [En ligne], 173 | Printemps 2006, mis en ligne le 22 mai 2006, consulté le 19 avril 2019. URL : http://journals.openedition.org/msh/2963 ; DOI : 10.4000/msh.2963

Math. & Sci. hum. ~ Mathematics and Social Sciences (44e année, n° 173, *2006*(1), p. 5-25)

COLINÉARITÉ ET RÉGRESSION LINÉAIRE

Thierry FOUCART[1]

RÉSUMÉ – *L'analyse linéaire de la régression, appelée aussi plus simplement régression linéaire, est l'une des méthodes statistiques les plus utilisées dans les sciences appliquées et dans les sciences de l'homme et de la société. Son objectif est double : il consiste tout d'abord à décrire les relations entre une variable privilégiée, appelée variable expliquée (ou dépendante), et plusieurs variables jouant un même rôle par rapport à la première, appelées variables explicatives (ou indépendantes). Elle permet aussi d'effectuer des prévisions de la variable expliquée en fonction des variables explicatives.*

Les liaisons entre les variables explicatives exercent une influence très importante sur l'efficacité de la méthode, quel que soit l'objectif dans lequel elle est utilisée. Nous exposons dans cet article des propriétés sur ces liaisons démontrées et publiées récemment dans plusieurs articles.

MOTS-CLÉS – Corrélation, Instabilité, Régression bornée, Régression sur composantes principales, Transitivité.

SUMMARY – Collinearity and linear regression analysis

The linear analysis of the regression, called also more simply linear regression, is one of the most used statistical methods in applied sciences and social sciences. Its objective is double: first of all it consists in describing the relations between a variable, called explained (or dependent) variable, and several variables, called explanatory (or independent) variables. It also makes it possible to conduct forecasts of the explained variable in terms of the explanatory variables. The links between the explanatory variables exert a considerable influence on the effectiveness of the method, whatever the objective in which it is used. We expose in this paper some of the properties of these links, recently proved and published in several papers.

KEYWORDS – Correlation, Instability, Principal Component Regression Analysis, Ridge Regression, Transitivity.

1. ANALYSE DE LA COLINÉARITÉ DANS LE MODÈLE LINÉAIRE

1.1 MODÈLE LINÉAIRE

Le modèle linéaire exprime mathématiquement la relation supposée entre une variable statistique, appelée variable expliquée (ou dépendante) et notée Y, et p variables

1 UMR 6086, Mathématiques SP2MI, bd Marie et Pierre Curie, BP 30179 – 86962 Futuroscope Chasseneuil cedex, foucart.thierry@free.fr

La plupart des méthodes présentées dans cet article ont été programmées. Le programme est disponible à l'adresse internet suivante : http://foucart.thierry.free.fr/

appelées variables explicatives (ou indépendantes) $X_1, \ldots, X_p$. On note n le nombre d'individus statistiques considérés, y_i la i^e observation de la variable Y et x_i^j celle de la variable X_j. Pour simplifier, nous supposons dans toute la suite que ces variables sont centrées et réduites.

$$\sum_{i=1}^{n} y_i = 0 \qquad \sum_{i=1}^{n} y_i^2 = n$$

$$\forall\ j = 1,\ldots, p \quad \sum_{i=1}^{n} x_i^j = 0 \quad \sum_{i=1}^{n} x_i^{j\,2} = n$$

Le modèle linéaire est défini par l'équation matricielle ci-dessous.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

dans laquelle.

- $\boldsymbol{Y}$ est le vecteur $(y_1, y_2, \ldots, y_n)^t$ des n valeurs observées de la variable expliquée Y ;
- $\boldsymbol{X}$ est la matrice des données à n lignes et p colonnes, la colonne j (de 1 à p) étant définie par le vecteur $(x_1^j, x_2^j, \ldots, x_n^j)^t$;
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ est le vecteur des coefficients de régression.
- $\boldsymbol{\varepsilon}$ est le vecteur résiduel $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^t$ défini par un échantillon indépendant de la variable résiduelle ε de variance σ^2.

On note $\boldsymbol{R}$ la matrice de corrélation entre les p variables $X_1, \ldots, X_p$. On la suppose inversible (de rang p). elle possède donc p valeurs propres strictement positives. La méthode classique d'estimation des paramètres est fondée sur le critère des moindres carrés. L'estimateur $\boldsymbol{B} = (B_1, B_2, \ldots, B_p)^t$ de $\boldsymbol{\beta}$ est alors donné par la formule ci-dessous.

$$\boldsymbol{B} = (1/n)\, \boldsymbol{R}^{-1}\, \boldsymbol{X}^t\, \boldsymbol{Y} = \boldsymbol{R}^{-1}\, \boldsymbol{r}$$

en notant $\boldsymbol{r}$ le vecteur des coefficients de corrélation observés $(r_1, r_2, \ldots, r_p)^t$ entre les variables explicatives X_j et la variable expliquée Y. Le vecteur $\boldsymbol{b} = (b_1, b_2, \ldots, b_p)^t$ est l'observation du vecteur $\boldsymbol{B}$. Le coefficient de détermination est par définition le carré du coefficient de corrélation linéaire de la variable expliquée $\boldsymbol{Y}$ et de $\boldsymbol{B}^t\boldsymbol{X}$. Il est égal à.

$$R^2 = \boldsymbol{r}^t\, \boldsymbol{R}^{-1}\, \boldsymbol{r}$$

La matrice variance de $\boldsymbol{B}$ est de la forme.

$$\boldsymbol{V_B} = \frac{\sigma^2}{n}\, \boldsymbol{R}^{-1}$$

On définit le vecteur des résidus $\boldsymbol{e} = (e_1, e_2, \ldots e_i, \ldots, e_n)^t$ en remplaçant le vecteur de régression $\boldsymbol{\beta}$ par son estimation $\boldsymbol{b}$.

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\, \boldsymbol{b}$$

D'après le théorème de Gauss-Markov, l'estimateur $\boldsymbol{B}$ est efficace (de variance minimale dans la classe des estimateurs linéaires sans biais).

1.2 COLINÉARITÉ ENTRE LES VARIABLES EXPLICATIVES

Les estimateurs précédents dépendent tous de la matrice $\boldsymbol{R}^{-1}$. Cette matrice n'existe pas si les variables $X_1, \ldots X_p$ sont colinéaires, ou, ce qui est équivalent, si une ou plusieurs valeurs propres de $\boldsymbol{R}$ sont nulles. Au plan numérique, il peut arriver qu'une liaison linéaire ne soit pas détectée, et que le calcul donne des résultats numériques aberrants. Cette difficulté est à peu près résolue depuis que les ordinateurs permettent une grande précision de calcul.

Au plan statistique, auquel nous nous limitons ici, l'existence d'une colinéarité approximative, appelée colinéarité statistique, peut perturber les estimations des paramètres du modèle. Une telle colinéarité peut exister, comme nous le précisons plus loin, même lorsque les coefficients de corrélation linéaire entre les variables $X_1, \ldots X_p$ sont faibles. Elle se manifeste par une ou plusieurs valeurs propres très petites de la matrice.

Les conséquences de la colinéarité statistique entre les variables explicatives sont les suivantes :

- les coefficients de régression estimés peuvent être élevés en valeur absolue ;
- leurs signes peuvent être contraires à l'intuition ;
- les variances des estimateurs peuvent être élevées ;
- les coefficients de régression et le coefficient de corrélation multiple sont instables par rapport aux coefficients de corrélation entre les variables explicatives.

La colinéarité statistique crée donc des difficultés importantes dans l'interprétation des résultats. Par exemple, le fait que le signe d'un coefficient de régression puisse être changé par la colinéarité peut être particulièrement gênant pour étudier l'effet propre d'une variable X_j sur Y.

On peut mesurer cette colinéarité de différentes façons. Hoerl et Kennard [1970(a)] montrent que l'erreur quadratique $E(\|\boldsymbol{B} - \beta\|^2)$ de $\boldsymbol{B}$ est égale à $\sigma^2 tr(\boldsymbol{R}^{-1})$ où $tr(\boldsymbol{R}^{-1})$ est la trace de la matrice $\boldsymbol{R}^{-1}$ et proposent comme mesure de la colinéarité les facteurs d'inflation f_j définis par les termes diagonaux de la matrice $\boldsymbol{R}^{-1}$. Ces termes diagonaux dépendent des coefficients de corrélation multiple R_j^2 obtenus en effectuant la régression de la variable X_j par les autres variables explicatives [Hawkins, Eplett, 1982] :

$$f_j = \frac{1}{1 - R_j^2}$$

Le facteur d'inflation f_j est donc d'autant plus grand que la variable X_j est corrélée à une combinaison linéaire des autres variables explicatives. Tomassone *et al.* [1992] appelle indice de multicolinéarité I la moyenne des facteurs d'inflation :

$$I = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{1 - R_j^2}$$

Considérons les valeurs propres $\lambda_1, \ldots, \lambda_p$ de la matrice $\boldsymbol{R}$ rangées dans l'ordre décroissant. Belsley, Kuh et Welsh [1980, p. 100] donnent la définition de l'indice de conditionnement κ et la façon de l'interpréter :

$$\kappa = \frac{1}{\lambda_p}$$

La matrice $\boldsymbol{R}^{-1}$ a pour valeurs propres les inverses des valeurs propres $\lambda_1, \ldots, \lambda_p$ précédentes. L'indice de multicolinéarité précédent s'exprime donc de la façon suivante :

$$I = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{\lambda_j}$$

Il tient compte de la présence éventuelle de plusieurs faibles valeurs propres et généralise l'indice de conditionnement.

Tomassone donne quelques explications sur les valeurs de l'indice de multicolinéarité *I*, mais pas de règle précise pour en apprécier la valeur. On remarquera que :

$$1 /(p\, \lambda_{p)} \leq I \leq 1 / \lambda_p$$

1.3 MÉTHODES DE RÉGRESSION DANS LE CAS DE COLINÉARITÉ

Il existe plusieurs estimateurs classiques des coefficients de régression limitant les difficultés dues à la colinéarité des variables explicatives.

- la régression des moindres carrés partiels [Helland, 1990 ; Tenenhaus, 1998 ; Wold *et al.*, 1984] : on obtient des estimateurs successifs en considérant les résidus comme une nouvelle variable dépendante ;
- les régressions pas à pas [Hocking, 1976] : en limitant le nombre de variables explicatives suivant leurs coefficients de corrélation partielle avec la variable expliquée, on réduit les colinéarités éventuelles ;
- la régression bornée, ou « ridge regression » [Hoerl, Kennard, 1970(b)]. On cherche un estimateur $\boldsymbol{B}_r$ des coefficients de régression sous la contrainte $\|\boldsymbol{B}_r\| < M$, M étant un réel positif fixé. Cet estimateur est donné par la formule ci-dessous :

$$\boldsymbol{B}_r = (1/n)\, [\, \boldsymbol{R} + k\, \boldsymbol{I}]^{-1}\, \boldsymbol{X}^{\,t}\, \boldsymbol{Y}$$

où k est un nombre réel positif choisi à l'aide de la représentation graphique des coefficients de régression en fonction de k, appelée « ridge trace ». Il existe une méthode donnant une valeur approximative de la meilleure valeur de k [Lee, 1988 ; Lee, Campbell, 1985 ; Nordberg, 1982 ; Tze-San, 1988] ;

- la régression orthogonale [Helland, 1992 ; Jolliffe, 1982 ; Naes, Helland, 1993 ; Rouanet *et al.* 2002 ; Webster *et al.* 1974]. Après avoir effectué l'analyse en composantes principales des variables $X_1, \ldots X_p$, on choisit comme variables explicatives les composantes principales C_l, $l = 1, \ldots p$. On élimine ensuite les composantes principales :
  - dont la corrélation avec la variable expliquée Y est faible, suivant un test de Fisher par exemple, pour réduire le nombre de variables explicatives ;
  - dont la variance, égale à la valeur propre correspondante, est faible. On considère en général que ces composantes principales sont instables et ne représentent qu'un bruit blanc.

La suppression de composantes principales revient à imposer une contrainte linéaire aux coefficients de régression, puisque ces composantes principales sont supposées égales à 0 sur tous les individus statistiques.

2. ANALYSE DESCRIPTIVE DE LA COLINÉARITÉ

2.1 DÉCOMPOSITION DE LA COLINÉARITÉ [Foucart, 1996]

Lorsque les variables explicatives X_j sont classées en plusieurs groupes G_l, $l = 1,\ldots, g$, on peut décomposer l'indice de multicolinéarité de façon analogue à la décomposition de la variance lorsque les individus statistiques sont classés en plusieurs groupes. Nous allons pour simplifier étudier le cas de deux groupes G_1 et G_2 de p_1 et p_2 variables ($p_1 + p_2 = p$). La matrice de corrélation $\boldsymbol{R}$ s'écrit de la façon suivante :

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{1,1} & \boldsymbol{R}_{1,2} \\ \boldsymbol{R}_{2,1} & \boldsymbol{R}_{2,2} \end{bmatrix}$$

où $\boldsymbol{R}_{1,1}$ et $\boldsymbol{R}_{2,2}$ sont les matrices de corrélation des variables des groupes G_1 et G_2, $\boldsymbol{R}_{1,2}$ la matrice de corrélation entre les variables du groupe G_1 et celles du groupe G_2 et $\boldsymbol{R}_{2,1}$ est égale à $\boldsymbol{R}_{1,2}{}^{t}$. Soient I_1 et I_2 les indices de multicolinéarité des variables dans G_1 et G_2. On a :

$$\frac{p_1 I_1 + p_2 I_2}{p} \leq I$$

D'où les définitions :

- l'indice de multicolinéarité totale est l'indice I calculé sur la totalité des variables ;
- l'indice de multicolinéarité résiduelle I_r est la moyenne des indices de multicolinéarité calculés sur les variables de chaque groupe :

$$I_r = \frac{p_1 I_1 + p_2 I_2}{p}$$

- l'indice de multicolinéarité expliquée I_e est la différence entre les deux précédents :

$$I_e = I - I_r \qquad \text{ou} \qquad I = I_e + I_r$$

On montre que l'indice de multicolinéarité expliquée est positif ou nul, et qu'il est nul si et seulement si les coefficients de corrélation de la matrice $\boldsymbol{R}_{1,2}$ sont tous nuls.

On définit le rapport de multicolinéarité μ^2 de la façon suivante :

$$\mu^2 = I_e / I$$

On a évidemment :

$$0 \leq \mu^2 \leq (I - 1) / I$$

Cette décomposition se généralise sans difficulté au cas d'un nombre quelconque de groupes de variables.

2.2 CLASSIFICATION DES VARIABLES

Le rapport de multicolinéarité μ^2 peut être utilisé comme critère de classification des variables. Le nombre de groupes étant choisi, on peut répartir les variables de façon à le maximiser. Cela revient à minimiser l'indice de multicolinéarité résiduelle. On obtiendra des groupes constitués de variables les moins liées possibles. Par contre ces groupes de variables seront liés canoniquement entre eux. On peut effectuer tous les calculs possibles dans le cas de deux groupes, et aboutir à un maximum absolu de μ^2. Dans le cas de plus de deux groupes, le calcul devient très long et on se limite à un extremum local.

La minimisation du rapport de multicolinéarité est aussi possible : on constitue des groupes dont les variables sont les plus colinéaires possibles. Cette classification peut être représentée sous la forme d'une arborescence : on regroupe, parmi les p groupes constitués chacun d'une variable initiale, les deux qui sont les plus corrélées, maximisant ainsi la multicolinéarité intra (ou minimisant le rapport de multicolinéarité). On obtient ainsi $p - 1$ groupes. Le regroupement suivant consiste à définir $p - 2$ groupes maximisant la multicolinéarité intra et ainsi de suite.

Exemple : les données que nous analysons ci-dessous sont extraites de l'ouvrage de Tomassone *et al*. [1992]. Le tableau est constitué de onze variables observées sur 33 unités statistiques, les dix premières étant les variables explicatives. Ces unités sont des parcelles forestières sur lesquelles on a observé :

X_1	altitude (en m)
X_2	pente (en degré)
X_3	nombre de pins moyens dans une placette de 5 ares
X_4	hauteur de l'arbre échantillonné au centre de la placette
X_5	diamètre de cet arbre
X_6	note de densité du peuplement
X_7	orientation de la placette (1 : sud, 2 : autres
X_8	hauteur (en m) des arbres dominants
X_9	nombre de strates de végétation
X_{10}	mélange du peuplement (1 : pas mélangé, 2 : mélangé)
Y	logarithme du nombre de nids de processionnaires par arbre d'une placette

La matrice de corrélation est la suivante :

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
X_1	1.000										
X_2	0.121	1.000									
X_3	0.538	0.322	1.000								
X_4	0.321	0.137	0.414	1.000							
X_5	0.284	0.113	0.295	0.905	1.000						
X_6	0.515	0.301	0.980	0.439	0.306	1.000					
X_7	0.268	-0.152	0.128	0.058	-0.079	0.151	1.000				
X_8	0.360	0.262	0.759	0.772	0.596	0.810	0.060	1.000			
X_9	0.364	0.326	0.877	0.460	0.267	0.909	0.063	0.854	1.000		
X_{10}	-0.100	0.129	0.206	-0.045	-0.025	0.130	0.138	0.054	0.175	1.000	
Y	-0.534	-0.429	-0.518	-0.425	-0.201	-0.528	-0.230	-0.541	-0.594	-0.063	1.000

Matrice de corrélation entre X_1, …, X_{10} et Y
(données de Tomassone).

L'existence de colinéarités est attestée par l'indice de multicolinéarité et la valeur des trois dernières valeurs propres de la matrice carrée constituée des dix premières lignes et colonnes :

$I = 15.8$	$\lambda_8 = 0.055440$	$\lambda_9 = 0.043032$	$\lambda_{10} = 0.009560$

La première classification des variables consiste à construire deux groupes de variables de façon à maximiser le rapport de multicolinéarité. On obtient les deux groupes ci-dessous :

Groupe	variables	Indice de multicolinéarité
G_1	X_1, X_3, X_5, X_7, X_8	$I_1 = 2.39$
G_2	X_2, X_4, X_6, X_9, X_{10}	$I_2 = 3.08$

La colinéarité entre les dix variables apparaît ici comme une colinéarité entre les deux groupes constitués chacun de variables faiblement colinéaires.

multicolinéarité totale	inter	intra	rapport
15.80	13.06	2.73	0.83

Décomposition de l'indice de multicolinéarité total

La colinéarité entre les deux groupes est aussi caractérisée par la valeur des coefficients de corrélation canonique :

coefficients de corrélation canonique		
rang 1	rang 2	rang 3
0.99305	0.93745	0.67683

Ces propriétés peuvent être généralisées à plus de deux groupes et détaillées par une analyse factorielle comme la méthode STATIS [Foucart, 1984]. Nous avons proposé d'effectuer une régression sélective des caractères canoniques pour obtenir un estimateur de β [Foucart, 1999(a)].

La deuxième classification, fondée sur la minimisation pas à pas du rapport de multicolinéarité, aboutit à l'arborescence ci-dessous□

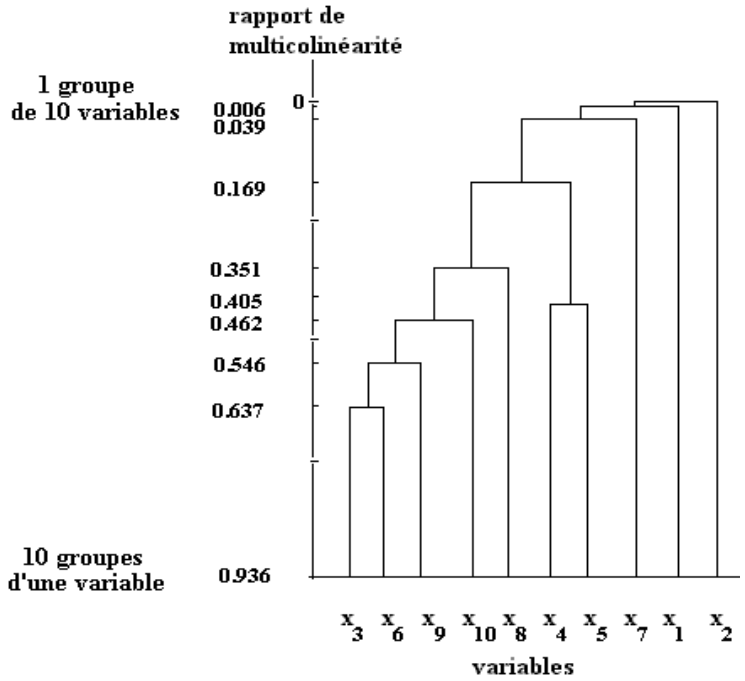

Figure 1□Classification des variables
par minimisation du rapport de multicolinéarité

Le rapport de multicolinéarité est maximal et égal à 0.936 lorsque les groupes sont constitués d'une seule variable, et à 0 lorsque toutes les variables ont été regroupées. Cette classification donne l'ordre dans lequel les variables se regroupent de façon à minimiser le rapport de multicolinéarité, ou à maximiser la multicolinéarité intra. Le groupe $\{X_3, X_6, X_9, X_{10}\}$ est constitué de variables fortement liées entre elles, de même que le groupe $\{X_4, X_5\}$. Les variables X_1, X_2 et X_7 ne sont guère liées aux précédentes ni entre elles.

3. PROPRIETES NUMERIQUES D'UNE MATRICE SYMETRIQUE DEFINIE POSITIVE

La régression consiste à raisonner conditionnellement aux valeurs observées des variables explicatives. L'analyse de la colinéarité peut donc être menée par l'étude numérique de la matrice $\boldsymbol{R}$, matrice symétrique définie positive particulière.

3.1 ENCADREMENT D'UN TERME D'UNE MATRICE SYMÉTRIQUE DÉFINIE POSITIVE [Foucart, 1991]

Soit $\boldsymbol{M} = (m_{i,j})$ une matrice symétrique définie positive de p lignes et p colonnes. On sait que toutes ses valeurs propres sont strictement positives et qu'elle est inversible. Le calcul de la matrice inverse est effectué par l'algorithme de Cholesky [Ciarlet, 1989□; Graybill, 1983□;Hawkins, Eplett, 1982].

PROPOSITION 1 (factorisation de Cholesky) : La matrice $\boldsymbol{M}$ est le produit d'une matrice triangulaire inférieure $\boldsymbol{T} = (t_{i,j})$ par sa transposée $\boldsymbol{T}^{t}$.

$$\boldsymbol{M} = \boldsymbol{T}\,\boldsymbol{T}^{t}$$

Cette factorisation permet d'énoncer la proposition suivante□

PROPOSITION 2 (généralisation de l'inégalité de Schwarz) : le terme $m_{p-1,p}$ de la matrice $\boldsymbol{M}$ appartient à l'intervalle] $a_{p-1,p}$, $b_{p-1,p}$ [défini par□

$$a_{p-1,p} = -\,t_{p-1,p-1}\,[m_{p,p} - \sum_{k=1}^{p-2} t_{p,k}^{\,2}]^{1/2} + \sum_{k=1}^{p-2} t_{p-1,k}\,t_{p,k}$$

$$b_{p-1,p} = t_{p-1,p-1}\,[m_{p,p} - \sum_{k=1}^{p-2} t_{p,k}^{\,2}]^{1/2} + \sum_{k=1}^{p-2} t_{p-1,k}\,t_{p,k}$$

Le terme $m_{p,p}$ est minoré par $c_{p,p}$ défini par□

$$c_{p,p} = \sum_{k=1}^{p-1} t_{p,k}^{\,2}$$

Les formules données ci-dessus concernent les termes $m_{p-1,p}$ et $m_{p,p}$□ On peut minorer et majorer tout terme de la matrice par simple permutation de lignes et de colonnes. On peut donc exprimer la propriété dans le cas général de la façon suivante□

PROPOSITION 3. Chaque terme d'une matrice symétrique définie positive appartient à un intervalle] *a, b* [appelé intervalle de variation, dont les bornes dépendent des autres termes, avec $b = +\infty$ si le terme est diagonal.

Exemple. Considérons la matrice symétrique ci-dessous□

$\boldsymbol{R} =$

	X	*Y*	*Y*
X	1		
Y	0.8	1	
Z	0.5	$r_{3,2}$	1

La matrice symétrique $\boldsymbol{R}$ est définie positive pour toute valeur de $r_{3,2}$ appartenant à l'intervalle de variation ci-dessous□

] *a, b* [=] -0.1196152 , 0.9196153 [

3.2. TRANSITIVITÉ DE LA CORRÉLATION

Cet encadrement résout complètement le problème classique de la transitivité de la corrélation. Dans le cas le plus simple, ce problème s'énonce de la façon suivante :

« *X* est fortement corrélée à *Y*, *Y* à *Z*. Peut-on en déduire que *X* et *Z* sont fortement corrélées ? »

Le mieux est de donner un exemple numérique. On considère la matrice ***R*** des corrélations entre les variables *X*, *Y* et *Z* ci-dessous :

$\boldsymbol{R} =$

	X	*Y*	*Y*
X	1		
Y	$r_{2,1}$	1	
Z	$r_{3,1}$	$r_{3,2}$	1

Le problème posé est le suivant : des valeurs élevées de $r_{1,2}$ (= $r_{2,1}$) et de $r_{2,3}$ (= $r_{3,2}$) impliquent-elles une forte valeur de $r_{1,3}$ (= $r_{3,1}$) ?

Supposons $r_{2,1}$ = 0.8. Sur la Figure 2 ci-dessous, la zone grise représente l'ensemble des couples ($r_{3,2}$, $r_{3,1}$) tels que la matrice ***R*** soit symétrique définie positive. Le centre des intervalles appartient à une droite.

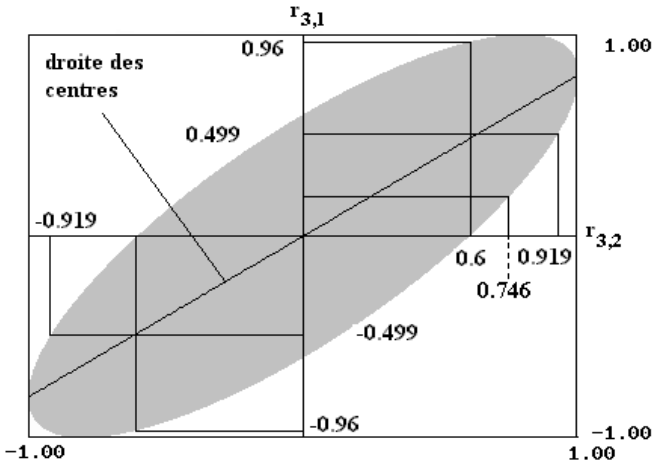

Figure 2. représentation graphique des couples ($r_{3,2}$, $r_{3,1}$)
tels que la matrice ***R*** soit symétrique définie positive pour $r_{2,1}$ = 0.8.

On a en particulier :

$$r_{2,3} > 0.600 \quad \Rightarrow \quad 0 < r_{1,3} < 0.960$$

Lorsque le coefficient de corrélation $r_{2,3}$ est supérieur à 0.6, le coefficient de corrélation $r_{1,3}$ est nécessairement positif : il y a donc transitivité de la corrélation, au plan numérique. On peut calculer facilement la valeur de $r_{2,3}$ telle que cette transitivité donne un coefficient $r_{1,3}$ statistiquement significatif ($r_{1,3} > 0.196$) pour un nombre d'observations égal à 100 :

$$r_{2,3} > 0.746 \quad \Rightarrow \quad r_{1,3} > 0.196$$

La généralisation de cette propriété à une matrice de taille quelconque ne pose pas de difficulté : on raisonne sous la condition que tous les coefficients de la matrice sont fixés sauf $r_{i,j}$ et $r_{j,k}$. On peut étudier aussi la relation entre deux coefficients de corrélation de la forme $r_{i,j}$ et $r_{k,l}$ avec $i \neq k$, $i \neq l$, $j \neq l$ et $j \neq k$. L'ensemble des couples $(r_{i,j}, r_{k,l})$ n'est plus nécessairement symétrique ni les centres des intervalles de variation alignés.

3.3 UNE AUTRE INTERPRÉTATION DU COEFFICIENT DE CORRÉLATION PARTIELLE

[Foucart, 1992, 1997(a)]

La transitivité de la corrélation est liée à la corrélation partielle. On démontre la propriété suivante :

PROPOSITION 4. Le coefficient de corrélation $r_{i,j}$ et le coefficient de corrélation partielle $rp_{i,j}$ vérifient la relation ci-dessous :

$$rp_{i,j} = \frac{r_{i,j} - (a+b)/2}{(b-a)/2}$$

Le coefficient de corrélation partielle $rp_{i,j}$ mesure la distance relative du coefficient de corrélation $r_{i,j}$ au centre de son intervalle de variation. La droite des centres représentée en figure 2 est l'ensemble des coefficients de corrélation $r_{3,1}$ tels que le coefficient de corrélation partielle $rp_{3,1}$ soit nul pour la valeur de $r_{3,2}$ considérée.

PROPOSITION 5. Les coefficients de corrélation $r_{k,l}$ étant donnés pour $(k, l) \neq (i, j)$, le coefficient de corrélation partielle $rp_{i,j}$ est une fonction linéaire croissante du coefficient de corrélation $r_{i,j}$, est égal à 0 lorsque $r_{i,j}$ est au centre de son intervalle de variation et tend vers 1 en valeur absolue lorsque $r_{i,j}$ tend vers l'une des bornes de cet intervalle.

Le coefficient directeur de cette droite est égal à $2/(b - a)$: il mesure la dépendance du coefficient de corrélation partielle par rapport au coefficient de corrélation et dépend des autres coefficients de corrélation.

Ces propriétés peuvent être généralisées au cas de tout coefficient de corrélation et de tout coefficient de corrélation partielle. Il suffit de remplacer les termes a et b par leurs expressions données dans le paragraphe 1.1 pour établir la relation entre le coefficient de corrélation partielle $rp_{i,j}$ et tout coefficient de corrélation $r_{k,l}$.

Les applications numériques montrent qu'en général, dès qu'un coefficient de corrélation tend vers une des bornes de son intervalle de variation, les autres aussi, et tous les coefficients de corrélation partielle tendent vers 1. Les cas particuliers sont ceux d'une matrice constituée comme dans le paragraphe 2.1, et dont une sous-matrice $\boldsymbol{R}_{1,2}$ est nulle. Ils correspondent à un rapport de multicolinéarité entre les groupes de variables G_1 et G_2 égal à 0.

4. ANALYSE DE LA COLINEARITE PAR DERIVATION

Les propriétés numériques précédentes permettent de dériver n'importe quel terme de la matrice inverse par rapport à n'importe quel terme de la matrice directe. On peut en déduire les dérivées de tous les paramètres de la régression : coefficient de

détermination, coefficients de régression, facteurs d'inflation, indice de multicolinéarité, variances des estimateurs… [Foucart, 1997(b), 2000(a)].

4.1 INSTABILITÉ DE L'INVERSE DE LA MATRICE DE CORRÉLATION

Considérons l'exemple numérique donné par la matrice de corrélation ci-dessous⊡

$\boldsymbol{R} =$

	X_1	X_2	X_3	Y
X_1	1			
X_2	0.6	1		
X_3	-0.279	0.5	1	
Y	0.0446	0	0	1

Les coefficients de corrélation ne présentent apparemment aucune particularité. On peut penser que la régression de Y par les variables X_1, X_2 et X_3 ne donnera pas de résultat significatif, compte tenu de la valeur des termes de la dernière ligne tous très proches de 0.

Contrairement à cette intuition, le coefficient de détermination obtenu dans la régression de Y par X_1, X_2, X_3 est très élevé⊡

$$R^2 = 0.99536$$

Par contre, en diminuant $r_{2,1}$ de 0.001 ($r_{2,1} = 0.599$ au lieu de $r_{2,1} = 0.6$), on trouve⊡

$$R^2 = 0.45260$$

Cette variation du coefficient de détermination de plus de 50 % s'explique par le fait que le coefficient $r_{2,1}$ est très proche de la borne supérieure de son intervalle de variation⊡

$$] a, b [=] - 0.9356329, 0.6008329 [$$

Sur cet exemple, la colinéarité entre les variables X_1, X_2 et X_3 exerce donc une influence considérable sur les résultats de la régression. Pour étudier cette instabilité, nous proposons d'étudier les dérivées du coefficient de détermination par rapport aux coefficients de corrélation⊡

$dR^2/dr_{2,1} = 1194.429$		
$dR^2/dr_{3,1} = -994.579$	$dR^2/dr_{3,2} = 1192.563$	
$dR^2/dr_{4,1} = 44.635$	$dR^2/dr_{4,2} = -53.520$	$dR^2/dr_{4,3} = 44.565$

Dérivées du coefficient de détermination R^2
par rapport aux termes extra diagonaux de $\boldsymbol{R}$.

Ces dérivées montrent la sensibilité du coefficient de détermination R^2 aux variations des coefficients de corrélation $r_{2,1}$, $r_{3,2}$ et $r_{3,1}$ entre les variables explicatives X_1, X_2, X_3, tandis qu'il l'est beaucoup moins aux variations des coefficients de corrélation $r_{4,1}$, $r_{4,2}$ et $r_{4,3}$ entre la variable expliquée Y et les variables explicatives X_1, X_2, X_3.

4.2 GÉNÉRALISATION. RÉGRESSION BORNÉE PARTIELLE [Foucart, 1998, 1999(b)]

On peut utiliser ces dérivées dans le cadre de la régression bornée pour choisir à la fois la constante k et les termes diagonaux auxquels elle sera ajoutée. Cette opération va perturber légèrement les vecteurs propres de la matrice $\boldsymbol{R}$, que la régression bornée habituelle laisse invariants. Nous allons considérer comme exemple les données que Tomassone *et al.* [1992] a analysées par la régression bornée habituelle.

La régression des moindres carrés ordinaires donne les résultats suivants□

Coefficients	Estimations	Ecarts-types	t de Student	Facteurs d'inflation
b_1	–0.45925	0.16135	–2.846	1.877
b_2	–0.31580	0.12848	–2.458	1.190
b_3	0.52043	0.76197	0.683	41.871
b_4	–1.08148	0.47135	–2.294	16.022
b_5	0.80054	0.36073	2.219	9.384
b_6	–0.20560	0.90263	–0.228	58.756
b_7	–0.03570	0.15122	–0.236	1.649
b_8	0.34205	0.44720	0.765	14.423
b_9	–0.58486	0.39353	–1.486	11.169
b_{10}	–0.08917	0.15130	–0.589	1.651

Coefficients de régression obtenus par régression des moindres carrés ordinaires
(données de Tomassone, $n = 33$, $R^2 = 0.69$)

Les coefficients de régression b_3 et b_6 sont particulièrement affectés par la colinéarité. Ils sont opposés malgré un coefficient de corrélation $r_{3,6}$ égal à 0.980, et leurs variances et leurs facteurs d'inflation sont élevés. Les autres facteurs d'inflation élevés concernent b_4, b_5, b_8 et b_9. Les coefficients b_4 et b_5 ($r_{4,5} = 0.905$) sont élevés en valeur absolue et opposés.

Le calcul des dérivées du carré de la norme du vecteur de régression et de l'indice de multicolinéarité par rapport aux termes diagonaux de la matrice $\boldsymbol{R}$ donne les résultats suivants□

	$/\, r_{1,1}$	$/\, r_{2,2}$	$/\, r_{3,3}$	$/\, r_{4,4}$	$/\, r_{5,5}$	$/\, r_{6,6}$	$/\, r_{7,7}$	$/\, r_{8,8}$	$/\, r_{9,9}$	$/\, r_{10,10}$
$d\,\|\|b\|\|^2$	– 2.43	– 0.46	– 47.63	– 73.82	– 33.62	– 23.00	0.43	– 17.84	– 0.34	– 1.00
dI	– 1.15	– 0.23	– 400.39	– 68.00	– 28.80	– 600.22	– 3.66	– 54.15	– 23.88	– 4.92

Dérivées du carré de la norme du vecteur de régression (1e ligne)
et de l'indice de multicolinéarité (2e ligne) par rapport aux termes diagonaux de $\boldsymbol{R}$

Un faible accroissement des termes diagonaux de rangs 3, 4, 5, 6, 8 et 9 de la matrice de corrélation fait donc fortement diminuer le carré de la norme du vecteur de régression et l'indice de multicolinéarité. Il n'est pas utile d'ajouter une constante à tous les termes diagonaux pour réduire l'effet de la colinéarité.

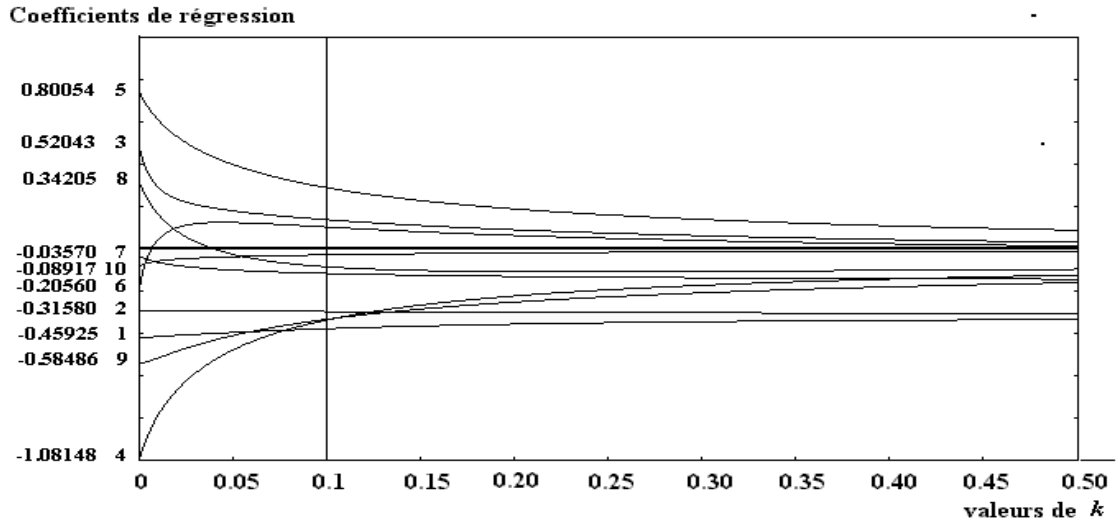

Figure 3. Représentation graphique des coefficients de régression en fonction de k (ridge trace, régression bornée partielle)

On choisit comme constante $k = 0.1$, à partir de laquelle les coefficients sont relativement bien stabilisés. Le carré de la norme du vecteur de régression est divisé par 4 (0.69 au lieu de 2.90) et l'indice de multicolinéarité est égal à 1.36 au lieu de 15.80. La régression bornée partielle a pour effet de diminuer fortement les écarts-types de b_3, b_4, b_5, b_6, b_8 et b_9.

$k = 0.1$	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}
b_j	– 0.409	– 0.321	0.149	– 0.364	0.312	0.111	– 0.126	– 0.092	– 0.360	– 0.031
s_j	0.159	0.137	0.144	0.149	0.150	0.124	0.139	0.166	0.167	0.136

Coefficients de régression obtenus en régression bornée partielle pour $k = 0.1$ (b_3, b_6, b_7, b_8, b_{10}, $R^2 = 0.65$)

Les coefficients de régression b_8 et b_{10} sont faibles par rapport aux écarts-types correspondants, et les variables X_8 et X_{10} peuvent être écartées des variables explicatives (on pourrait aussi envisager d'éliminer X_3, X_6, X_7). Dans le cas de la régression bornée habituelle, Hoerl et Kennard [1970(b)] recommandent de ne pas recommencer les calculs et de conserver tels quels les autres coefficients. Le modèle considéré est finalement le suivant□

$$Y = -0.409\,X_1 - 0.321\,X_2 + 0.149\,X_3 - 0.364\,X_4 + 0.312\,X_5 + 0.111\,X_6 - 0.126\,X_7 - 0.360\,X_9 + \varepsilon$$

5. REGRESSION ORTHOGONALE. CHOIX DES COMPOSANTES PRINCIPALES

5.1 RÉGRESSION ORTHOGONALE

La sélection d'une composante principale C_l comme variable explicative dépend de son coefficient de régression γ_l avec la variable expliquée Y – on cherche à limiter le nombre de variables explicatives – et de sa variance égale à la valeur propre associée λ_l – on cherche à éliminer les colinéarités.

Les variables explicatives peuvent être indifféremment les composantes principales C_l de variance λ_l ou les composantes principales réduites C_l'. Dans ce dernier cas, le modèle linéaire est défini par l'équation ci-dessous :

$$\boldsymbol{Y} = \boldsymbol{C'}\,\rho + \varepsilon$$

dans laquelle :

- $\boldsymbol{C'}$ est le tableau des composantes principales à n lignes et p colonnes, la colonne l (de 1 à p) étant définie par le vecteur $(c_1^l,\ c_2^l,\ ...,\ c_n^l)^t$ des coordonnées des unités statistiques sur l'axe principal de rang l ;
- $\rho = (\rho_1,\ ...,\ \rho_p)^t$ est le vecteur des coefficients de corrélation théoriques entre la variable expliquée Y et les composantes principales C_l des variables explicatives X_j, $j = 1, \ldots p$.

L'estimateur $\boldsymbol{B}_o$ des coefficients de régression sur les variables initiales est différent de l'estimateur des moindres carrés $\boldsymbol{B}$ lorsque certaines composantes principales ont été écartées. L'indice de multicolinéarité I_C est égal à :

$$I_C = \frac{1}{q} \sum_{j=i_1}^{i_q} \frac{1}{\lambda_j}$$

$i_1,\ i_2,\ \ldots i_q$ étant les rangs des composantes principales retenues. Il est évidemment inférieur à I si les composantes principales exclues sont associées aux valeurs propres les plus faibles.

La procédure classique, pour sélectionner une variable comme variable explicative, consiste à effectuer un test de Fisher Snedecor sur le coefficient de corrélation partielle. Suivant que la valeur observée du F appartient ou non à la région critique, on retient ou on exclut la variable considérée. Pour un risque de première espèce α fixé, on ne connaît pas le risque de seconde espèce β. En diminuant α, on augmente β, et le problème posé est donc d'optimiser le choix du risque α.

5.2 AUGMENTATION DE LA MOYENNE DES CARRÉS DES RÉSIDUS [Foucart, 2000(b)]

L'augmentation de la moyenne des carrés des résidus résulte des deux erreurs éventuelles.

- La première consiste à retenir une composante principale C_l' alors que son coefficient de corrélation partielle vrai avec la variable expliquée est nul. Si l'on note r_l le coefficient de corrélation observé, l'augmentation des carrés des résidus due à cette erreur est égale à :

$$\Delta\, MSR1 = r_l^2$$

On montre que son espérance, pour α fixé, dépend du coefficient de détermination R^2_{-l} de Y et des composantes principales déjà introduites C_k, pour $k \neq l$:

$$\Delta\, MSR1(\alpha) = \alpha\,(1 - R^2_{-l}) / (n - p - 1)$$

- La seconde erreur consiste à rejeter une composante principale C_l' alors que son coefficient de corrélation partielle vrai est non nul. La probabilité de cette erreur

dépend de la valeur du coefficient de corrélation partielle ρp_l supposé dans l'hypothèse alternative et est notée $\beta(\rho p_l)$. Pour évaluer le risque de seconde espèce $\beta(\rho p_l)$, ρp_l étant fixé, on utilise l'approximation normale de la loi du coefficient donnée par la v.a. z de Fisher [Kendall, Stuart, 1961].

Il est ensuite nécessaire de supposer que le coefficient de corrélation partielle ρp_l suit une loi de probabilité dont la densité est définie *a priori* par une fonction *f*. Il s'agit ici d'une démarche bayesienne. L'espérance de l'erreur que l'on commet en rejetant la composante principale est égale à⊡

$$\Delta MSR2(\alpha) = (1 - R^2_{-l}) \int \rho_{pl}{}^2 \beta(\rho p_l) f(\rho p_l) \, d\rho p_l$$

La loi de probabilité *a priori* de ρp_l peut être approchée par la loi normale d'espérance égale au coefficient de corrélation partielle observé rp_l, de variance $1/n$ et tronquée entre 0 et 1. Les calculs montrent en effet la robustesse de cette méthode par rapport à la loi choisie.

L'erreur a finalement pour moyenne⊡

$$\Delta MSR(\alpha) = \Delta MSR1(\alpha) + \Delta MSR2(\alpha)$$

On cherche ensuite le risque de première espèce α_{min} pour lequel la somme $\Delta MSR(\alpha)$ est minimale. La recherche est numérique⊡on fait varier α de 0 à 1, et on en déduit la région critique du coefficient de corrélation partielle suivant le coefficient de détermination R^2_{-l}, ou de façon équivalente, suivant le coefficient de corrélation observé entre *Y* et C_l, et le coefficient de détermination R^2.

5.3 APPLICATION AUX DONNÉES DE TOMASSONE [Tomassone *et al.*, 1992]

La régression orthogonale des données de Tomassone donne les résultats suivants:

Composante principale	Valeur propre	Corrélation	Composante principale	Valeur propre	Corrélation
C_1	4.679	– 0.621	C_6	0.543	0.087
C_2	1.533	0.139	C_7	0.160	0.351
C_3	1.232	– 0.063	C_8	0.055	0.075
C_4	0.993	– 0.104	C_9	0.043	– 0.224
C_5	0.752	– 0.286	C_{10}	0.010	0.078

Corrélations Composantes principales *x* Variable expliquée *Y*
(données de Tomassone, $n = 33$)

On peut sélectionner les composantes principales⊡

- soit par un algorithme ascendant⊡aucune composante principale n'est sélectionnée *a priori*, et on les choisit au fur et à mesure⊡
- soit par un algorithme descendant⊡on effectue la régression en considérant toutes les composantes principales comme variables explicatives, et on procède par élimination.

Ces deux algorithmes diffèrent par les degrés de liberté et par les coefficients de détermination considérés R^2_{-l}.

Nous utilisons ci-dessous l'algorithme descendant. L'élimination des composantes principales C_3, C_4, C_6, C_8 et C_{10} est assez évidente : il reste donc 5 composantes principales.

Le coefficient de détermination est alors égal à $R^2 = 0.6611$. Pour décider si les composantes principales C_2 et C_9 doivent être conservées, on calcule :

- les coefficients de détermination R^2_{-2} et R^2_{-9} sans chacune de ces composantes principales :

$$R^2_{-2} = R^2 - 0.139^2 = 0.6418 \qquad R^2_{-9} = R^2 - 0.224^2 = 0.6109$$

- les coefficients de corrélation partielle de C_2 et C_9 avec Y :

$$rp_2 = r_2 / (1 - 0.6418)^{1/2} = 0.232 \qquad rp_9 = r_9 / (1 - 0.6109)^{1/2} = 0.359$$

L'algorithme décrit dans le paragraphe 5.2 donne comme risque de première espèce minimisant $DMSR(\alpha)$ 0.3353 pour C_2 et 0.3545 pour C_9, et les valeurs limites correspondantes des coefficients de corrélation 0.1115 et 0.1221. Les coefficients de corrélation de ces composantes principales avec Y sont donc significatifs et il y a lieu de conserver C_2 et C_9.

Les coefficients de régression obtenus sur les variables initiales centrées réduites sont donnés ci-dessous :

C_1, C_2, C_5, C_7, C_9	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
b_j	– 0.455	– 0.293	– 0.012	– 0.963	0.772	0.361	0.015	0.042	– 0.398	– 0.195
s_j	0.107	0.098	0.217	0.387	0.277	0.128	0.061	0.261	0.115	0.050
t_j	– 4.26	– 3.01	– 0.06	– 2.49	2.79	2.83	0.24	0.16	– 3.46	– 3.90

Coefficients de régression obtenus en régression orthogonale
(composantes principales C_1, C_2, C_5, C_7, C_9, $R^2 = 0.66$)

Ces résultats méritent quelques commentaires :

- les composantes principales retenues sont celles qui minimisent l'estimation sans biais de la variance résiduelle et la statistique de Mallows [1973] ;
- on peut éliminer X_3, X_7 et X_8 des variables explicatives ;
- la sélection de C_9 introduit une colinéarité entre les variables explicatives compte tenu de sa faible variance ($\lambda_9 = 0.043$).

En éliminant C_9, on obtient les résultats suivants :

C_1, C_2, C_5, C_7	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
b_j	– 0.497	– 0.304	0.317	– 0.200	0.275	0.244	– 0.060	– 0.416	– 0.382	– 0.196
s_j	0.110	0.102	0.148	0.069	0.129	0.119	0.051	0.131	0.121	0.053
t_j	– 4.50	– 2.96	2.13	– 2. 90	2.13	2.04	– 1.18	– 3.19	– 3.16	– 3.72

Coefficients de régression obtenus en régression orthogonale
(composantes principales C_1, C_2, C_5, C_7, $R^2 = 0.61$)

L'élimination de C_9 donne des coefficients de régression b_3 et b_6 de même signe. Le modèle de régression est toutefois assez différent de celui qui est donné par la régression bornée partielle, et le coefficient de détermination est nettement diminué par cette élimination (0.61 au lieu de 0.66).

6. DISCUSSION

Les résultats précédents permettent d'approfondir l'interprétation des résultats d'une régression linéaire multiple. L'estimateur des moindres carrés ordinaires présente des inconvénients dans le cas de variables explicatives statistiquement colinéaires, et la recherche d'un estimateur satisfaisant n'est guère facile. On sait correctement déceler la présence d'une colinéarité, mais les méthodes adaptées donnent, comme on a pu le constater, des coefficients de régression assez différents les uns des autres, alors que les coefficients de détermination entre la variable estimée et la variable expliquée et calculés sur le fichier de donnés sont très voisins (de 0.6 à 0.7).

Il est bien difficile de choisir entre les modèles ainsi obtenus. On pourrait certes comparer leurs résultats sur un fichier test, mais cela n'apporte qu'une information très partielle, puisque ce fichier test n'est pas en général structuré comme le fichier de calcul lorsqu'il existe une composante principale de faible variance et par suite instable, et que l'écart entre les résultats obtenus par deux modèles n'est pas toujours suffisant pour lever l'ambiguïté.

L'utilisateur peut rechercher des coefficients de régression particuliers. Sur les données de Tomassone, on peut vouloir des coefficients b_3 et b_6 proches l'un de l'autre, compte tenu du coefficient de corrélation $r_{3,6}$ ($r_{3,6} = 0.980$). La régression bornée montre que c'est possible, et que les effets propres de X_3 et X_6 sur Y peuvent être considérés comme les mêmes malgré les estimations b_3 et b_6 de signes opposés données par les moindres carrés ordinaires. Par contre, aucune des méthodes précédentes ne permet d'obtenir des estimations b_4 et b_5 de même signe malgré un coefficient de corrélation $r_{4,5}$ élevé ($r_{4,5} = 0.905$): l'effet propre de X_4 sur Y semble donc l'opposé de celui de X_5. Cette notion d'effet propre est discutable, dans la mesure où il est bien difficile d'étudier l'effet sur Y d'une variation de X_4 (ou X_3) en supposant X_5 (ou X_6) constante, compte tenu de leur coefficient de corrélation.

Lorsque l'objectif de la régression est de décrire les relations entre les variables observées, on raisonne en supposant les variables explicatives fixées, et l'algorithme de sélection des composantes principales proposé semble satisfaisant. On garderait donc C_1, C_2, C_5, C_7, C_9 dans la régression effectuée sur les données de Tomassone, en privilégiant le coefficient de corrélation de C_9 avec la variable expliquée par rapport à la valeur propre λ_9. Par contre, pour effectuer une prévision, on recherche des coefficients

de régression plus stables, que l'on obtient en éliminant C_9 des composantes principales précédemment retenues. Pour cela, on peut introduire une fonction de coût liée à la valeur propre dans le calcul du risque moyen $\Delta\, MSR(\alpha)$.

Cette procédure de sélection des composantes principales donne des indications sur le risque de première espèce à choisir. Ce risque dépend à la fois de la taille de l'échantillon, du coefficient de détermination avec les composantes principales, et du coefficient de corrélation de la composante principale considérée et de la variable expliquée. En général, il est de l'ordre de 0.3□ il reste à savoir si c'est une valeur correcte dans les régressions pas à pas classiques, stepwise, ascendante ou descendante.

Il y a deux inconvénients à l'algorithme□

- le résultat est indépendant de la variance de la composante principale□;
- il ne permet pas d'établir une table.

On notera que la régression bornée est très efficace dans le cas de variables explicatives colinéaires, au point que l'on a intérêt à l'utiliser systématiquement en choisissant une faible valeur de la constante k. Cette efficacité est montrée par des procédures de simulation que l'on trouvera dans un article en cours de publication [Foucart, 2006].

Dans un document interne enfin, on recherche les valeurs influentes dans la régression en dérivant les estimations des coefficients de régression et du coefficient de détermination par rapport aux valeurs des variables observées x_j^i.

Remerciements. Je remercie ici les deux referees qui ont relu ce texte avec attention.

BIBLIOGRAPHIE

BELSLEY D.A., KUH E., WELSH R.E., *Regression diagnostics: identifying influential data and sources of collinearity*, New York, Wiley, 1980.

CIARLET P.G., *Introduction to numerical linear algebra and optimisation*, London, Cambridge University Press, 1989.

FOUCART T., *Analyse factorielle de tableaux multiples*, Paris, Masson, 1984.

FOUCART T., «□Transitivité du produit scalaire□», *Rev. Statistique Appliquée*, XXXIX (3), 1991, p. 57-68.

FOUCART T., «□Colinéarité dans une matrice de produit scalaire□», *Rev. Statistique Appliquée*, XXXX (3), 1992, p. 5-17.

FOUCART T., «□Analyse de la colinéarité. Classification de variables□», *Rev. Statistique Appliquée*, XXXX (3), 1996, p. 5-17.

FOUCART T., «□Numerical analysis of a correlation matrix□», *Statistics*, 29/4, 1997(a), p. 347-361.

FOUCART T., «□Stabilité de la matrice inverse d'une matrice symétrique définie positive□», *Comptes Rendus de l'Académie des Sciences*, Paris, t. 325, Série I, 1997(b), p. 91-96.

FOUCART T., « Régression bornée partielle », *Comptes Rendus de l'Académie des Sciences*, Paris, t. 326, Série I, 1998, p. 759-762.

FOUCART T., « Linear multiple regression on canonical variables », *Biometrical Journal* 41(5), 1999(a), p. 559-572.

FOUCART T., « Stability of the inverse correlation matrix. Partial ridge regression », *Journal of statistical planning and inference*, n° 77, 1999(b), p. 141-154.

FOUCART T., « Colinéarité et instabilité numérique dans le modèle linéaire », *RAIRO Operations research*, Vol. 34(2)2, 2000(a) p. 199-212.

FOUCART T., « A decision rule to discard principal components in regression », *Journal of statistical planning and inference*, n° 89, 2000(b), p. 187-195.

FOUCART T., « Sur l'efficacité de la régression bornée », *Revue des nouvelles technologies de l'information*, Cépaduès éd., 2006 [à paraître].

GRAYBILL F.A., *Matrices with applications in statistics*, Belmont (California), Wadsworth International Group, 1983.

HAWKINS D.M., EPLETT W.J.R., « The cholesky factorization of the inverse correlation or covariance matrix in multiple regression », *Technometrics* 24, 1982, p. 191-198

HELLAND I.S., « Maximum Likelihood Regression on Relevant Components », *J. R. Statist. Soc.* B, 54, n° 2, 1992, p. 637-647.

HELLAND, I.S., « Partial least squares regression and statistical models », *Scand. J. Statist.* 17, 1990, p. 97-114

HOCKING, R.R., « The analysis and selection of variables in linear regression », *Biometrics* 32, 1976, p. 1-40.

HOERL A.E., KENNARD R.W., « Ridge regression: biased estimation for nonorthogonal problems », *Technometrics* 12, 1970(a), p. 55-67.

HOERL A.E., KENNARD R.W., « Ridge regression: applications to nonorthogonal problems », *Technometrics* 12, 1970(b), p. 69-82.

JOLLIFFE I.T., « A note on the use of principal components in regression », *Appl. Statist.* 31, 1982, p. 300-303.

KENDALL M.G., STUART, A., *The advanced theory of statistics*, vol. 2**,** *Inference and relationship*, Griffin, London, 1961.

LEE T.S., « Optimum ridge parameter selection », *Appl. Statist.* 36(1), 1988, p. 112-118

LEE T.S., CAMPBELL D.B., « Selecting the optimum k in ridge regression », *Commun. Statist.*, A14, 1985, p. 1589-1604.

MALLOWS C.L., « Some comments on C_p »; *Technometrics* 15, 1973, p. 661-676.

NAES T., HELLAND I.S., « Relevant components in regression » *Scand. J. Statis.* 20, 1993, p. 239-250.

NORDBERG L., « A procedure of determination of a good ridge parameter in linear regression. Commun », *Statist. Simula. Computa.* 11(3), 1982, p. 285-289.

ROUANET H., LEBARON F., LE HAY V., ACKERMAN W., LE ROUX B., « Régression et analyse géométrique des données : réflexions et suggestions », *Mathématiques et Sciences humaines* 160, 2002, p. 13-45.

TENENHAUS M., « La régression PLS, théorie et pratique », *Technip*, Paris, 1998.

TOMASSONE R., AUDRAIN S., LESQUOY DE TURCKHEIM E.,MILLIER C., *La régression, nouveaux regards sur une ancienne méthode statistique*, 2e édition, Paris, Masson, 1992.

TZE-SAN, L., «Optimum ridge parameter selection», *Appl. Statist.* 36(1), 1988, p. 112-118.

WEBSTER J.T., GUNST R.F., MASON R.L., «Latent root regression analysis», *Technometrics* 16, n° 4, 1974, p. 513-522.

WOLD S., RUHE A., WOLD H., DUNN III W.J., «The collinearity problem in linear regression. The partial least squares (PLS) approached to generalized inverses», *SIAM Sci .Stat. Comp.*, Vol 5, n° 3, 1984, p. 735-743.